

# SEMI-INFINITE PROGRAMMING USING HIGH-DEGREE POLYNOMIAL INTERPOLANTS AND SEMIDEFINITE PROGRAMMING

DÁVID PAPP

**ABSTRACT.** In a common formulation of semi-infinite programs, the infinite constraint set is a requirement that a function parametrized by the decision variables is nonnegative over an interval. If this function is sufficiently closely approximable by a polynomial or a rational function, then the semi-infinite program can be reformulated as an equivalent semidefinite program. Solving this semidefinite program is challenging if the polynomials involved are of high degree, due to numerical difficulties and bad scaling arising both from the polynomial approximations and from the fact that the semidefinite programming constraints coming from the sum-of-squares representation of nonnegative polynomials are badly scaled. We combine rational function approximation techniques and polynomial programming to overcome these numerical difficulties, using sum-of-squares interpolants. Specifically, it is shown that the conditioning of the reformulations using sum-of-squares interpolants does not deteriorate with increasing degrees, and problems involving sum-of-squares interpolants of hundreds of degrees can be handled without difficulty. The proposed reformulations are sufficiently well scaled that they can be solved easily with every commonly used semidefinite programming solver, such as SeDuMi, SDPT3, and CSDP. Motivating applications include convex optimization problems with semi-infinite constraints and semidefinite conic inequalities, such as those arising in the optimal design of experiments. Numerical results align with the theoretical predictions; in the problems considered, available memory was the only factor limiting the degrees of polynomials, to approximately 1000.

## 1. INTRODUCTION

A linearly constrained semi-infinite convex optimization problem (SILP) with infinitely many convex constraints can be posed as:

$$\begin{aligned} & \text{minimize}_x && f(x) \\ & \text{subject to} && A(t)x \leq b \quad \forall t \in T \\ & && x \in X \end{aligned} \tag{SILP}$$

with respect to the decision variables  $x$ , where the set  $X \subseteq \mathbb{R}^n$  is convex, closed and bounded,  $f$  is convex on  $X$ , and the set  $T$  indexing the constraints is an arbitrary (usually low-dimensional, but not necessarily convex) uncountable set. Even without any restrictions on the dependence of  $A$  on  $t$ , (SILP) is a convex optimization problem, and the Weierstrass extreme value theorem guarantees that its minimum is attained.

In many applications,  $x$  represents a function  $p: T \mapsto \mathbb{R}$  that is known to belong to a given finite dimensional linear space (that is,  $x$  is the coefficient vector of  $p$  in some fixed basis), and the infinite constraint set represents  $p(t) \geq 0$  for all  $t \in T$ . Similar constraints on the derivatives of a differentiable function, such as  $\frac{dp(t)}{dt} \geq 0$  (implying monotonicity) or

$\frac{d^2 p(t)}{dt^2} \geq 0$  (implying convexity) can also be represented in a similar fashion. Optimization models incorporating such constraints have been used, for example, in arrival rate estimation [1], and in semi-parametric density estimation with and without shape constraints [29].

In most applications, including all the ones mentioned above, the index set  $T$  is low-dimensional, and is often simply an interval or a two-dimensional rectangular box [23, 16, 33].

Several algorithms have been proposed to solve semi-infinite linear and semi-infinite convex programming problems, including cutting plane and cutting surface methods such as [20, 5, 25], local reduction methods [16], exchange methods [40], and homotopy methods [21]. See also [23] for a relatively recent review on semi-infinite convex programming, including an overview on numerical methods with plenty of references.

The primary motivation behind this work is the case when solutions need to be computed with high accuracy, which is often the case when the results of the optimization are inputs of further numerical computations that are sensitive to errors in their inputs. Such problems arise for example in the computation of optimal designs of experiments, where the points in  $T$  for which the semi-infinite constraints are active need to be determined. We shall revisit this problem, and discuss it in more detail, in Section 6.3.

The methods mentioned above for solving (SILP) rarely have faster than linear convergence, which makes them particularly challenging to apply when the solutions are needed to be computed with high accuracy. On the other hand, in special cases that admit “nice” finitely constrained convex optimization formulations, or a linear cone optimization formulation over “nice” cones such as symmetric cones, superlinearly convergent (or even polynomial time) interior point methods can be utilized to compute solutions with high accuracy. These special cases include when all the sets and functions involved in the problem formulation are *semidefinite representable* (see below), which implies that (SILP) can be posed as a semidefinite optimization problem.

Let  $\mathbb{S}^k$  denote the set of  $k \times k$  real symmetric matrices, and  $\mathbb{S}_+^k$  denote the set of positive semidefinite matrices from  $\mathbb{S}^k$ . We shall also use the notation  $M \succcurlyeq 0$  to denote that the matrix  $M$  is positive semidefinite. Recall [39], [3, Sec. 4.2] that a set  $S \subseteq \mathbb{R}^n$  is *semidefinite representable* if for some  $k \geq 1$  and  $\ell \geq 0$  there exist affine functions  $A : \mathbb{R}^n \mapsto \mathbb{S}^k$  and  $C : \mathbb{R}^\ell \mapsto \mathbb{S}^k$  such that the set  $S$  can be characterized by a linear matrix inequality in the following way:

$$S = \{s \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^\ell : A(s) + C(u) \succcurlyeq 0\}.$$

With this definition, we can also say that a continuous *function* is *semidefinite representable* if all of its closed lower level sets are semidefinite representable. (For maximization problems, a concave objective function is semidefinite representable if its closed upper level sets are semidefinite representable.)

For the rest of the paper we assume that

- (1) the objective function  $f$  is semidefinite representable;
- (2) the set  $X$  is semidefinite representable;
- (3) the set  $T$  is the union of finitely many closed and bounded intervals; and
- (4) the elements of  $A(t)$  are continuous on  $T$ .

These assumptions are practically nonrestrictive, and the first three of them imply that aside from the infinite constraint set  $A(t)x \leq b$ , the problem (SILP) can be posed as a semidefinite program. If, additionally,  $A(t)$  belongs to a “nice” set of univariate functions, for which the set  $\{x \in \mathbb{R}^n \mid A(t)x \leq b \ \forall t \in [a, b]\}$  is also semidefinite representable for every given interval  $[a, b]$ , then (SILP) can be reformulated as a semidefinite program, and can in principle be solved to high accuracy with existing semidefinite programming solvers.

The last two assumptions imply that the elements of  $A(t)$  can be approximated arbitrarily closely in the uniform norm by polynomials [35, Chapter 1]. Since the set of nonnegative polynomials is also semidefinite representable [19, 3], this suggests the following approach:

- (1) Replace all entries of  $A(t)$  by a polynomial or rational function that approximates the entry uniformly within a chosen  $\varepsilon > 0$  in the uniform norm, and
- (2) reformulate the resulting approximate optimization problem as a semidefinite program, and solve it using interior point methods.

While this approach is fairly “obvious”, it can easily be viewed as entirely impractical, not the least because the semidefinite representation of nonnegative polynomials (when the polynomials are represented in the standard basis) involves Hankel matrices that are notoriously ill-conditioned [38, 2], and existing interior point methods are not designed to be able to handle such problems numerically. Although a change of basis to some orthogonal basis, such as the Chebyshev or Legendre polynomial basis, can somewhat mitigate this problem, the resulting semidefinite programs can still be difficult, if not impossible, to solve with existing semidefinite programming solvers, once the degree of the polynomials involved exceeds about 40 [27]. In contrast, a very close approximation of  $A(t)$  may require polynomials with at least hundreds of degrees.

As the main contribution of this paper, we demonstrate that using a suitable problem-dependent basis of polynomial interpolants to represent the approximators of  $A(t)$ , the above approach of polynomial approximation and semidefinite reformulation can be carried out efficiently, completely circumventing the aforementioned numerical difficulties and scaling issues, even if the polynomials involved are of high degree.

From an implementation perspective, this means combining efficient and numerically stable algorithms for manipulating polynomial interpolants and standard semidefinite programming solvers. Beyond its simplicity, an additional advantage of this approach over developing tailor-made algorithms to handle the polynomial nonnegativity constraints is that this approach is directly applicable to convex optimization problems that involve semidefinite constraints other than the ones arising from polynomial nonnegativity. A family of such problems is presented in Section 6.3.

The rest of the paper is structured as follows. In Section 2, we briefly review existing results regarding best polynomial approximations of smooth functions, and their representations as Lagrange or Hermite interpolants. In Section 3, we extend the results of [22] and [10] to provide semidefinite representations of sum-of-squares interpolants. Following that, we discuss how sum-of-squares Lagrange interpolants admit well-scaled semidefinite representations that can be handled with any existing semidefinite programming solver, in Section 4. Section 5 is concerned with a formally minor, but numerically critical issue of representing low-degree polynomials as high-degree interpolants in a fashion that does not affect the numerical properties of the semidefinite representation. This is an essential ingredient of optimization problems involving sum-of-squares polynomials of different degrees. Numerical examples are presented in Section 6. The first couple of examples illustrate the individual components developed throughout Sections 2–5, while the statistical applications (concerning optimal designs of experiments) demonstrate how the developed methodology can be used to solve convex semi-infinite programs with both functional and conic constraints using high-degree polynomial approximations and semidefinite programming. Section 7 concludes the paper with a discussion on the advantages and limitations of the approach.

## 2. TWO- AND ONE-SIDED APPROXIMATIONS

In order to keep the paper self contained, we shall very briefly review a few key results from constructive approximation theory that we shall make use of. The reader will find much more detail on these ideas in the references provided with the Propositions below. The next section connects these results to conic programming representations of best and near-best polynomial approximations of smooth functions.

**2.1. Lagrange interpolants and near-optimal two-sided approximation.** Consider a differentiable function  $f: [-1, 1] \mapsto \mathbb{R}$  that we wish to uniformly approximate by a polynomial, and take a set of distinct points  $t_0, \dots, t_n$  in the interval  $[-1, 1]$ . Then there is a unique polynomial  $p_n$  of degree  $n$  satisfying  $p(t_j) = f(t_j)$  for each  $j = 0, \dots, n$ . We call  $p_n$  the *polynomial (Lagrange) interpolant* of  $f$  on the *interpolation points*  $t_0, \dots, t_n$ . If the interpolation points are chosen well, increasing the number of interpolation points decreases the *interpolation error*  $\|f - p_n\| \stackrel{\text{def}}{=} \max_{-1 \leq t \leq 1} |f(t) - p_n(t)|$ . A particularly useful choice of interpolation points is the set of *Chebyshev points of the second kind*, defined by the formula

$$t_\ell = \cos(\ell\pi/n) \quad \ell = 0, \dots, n. \quad (1)$$

Another frequently used interpolation points are the *Chebyshev points of the first kind*; they are defined by the formula

$$t_\ell = \cos((\ell + 1/2)\pi/(n + 1)) \quad \ell = 0, \dots, n. \quad (2)$$

Many other schemes exist for choosing interpolation points.

Using the Chebyshev points defined in (1), the interpolation error converges to zero at a rate dependent on an appropriate measure of smoothness of  $f$ :

**Proposition 1** ([37, Theorem 7.2]). *For an integer  $k \geq 1$ , let  $f$  and its derivatives through  $f^{(k-1)}$  be absolutely continuous on  $[-1, 1]$ , and suppose that the  $k$ th derivative  $f^{(k)}$  is of bounded variation  $V_{f^{(k)}}$ . Then for every  $n \geq k + 1$ , the interpolant  $p_n$  through the Chebyshev points (1) satisfies*

$$\|f - p_n\| \leq \frac{4V_{f^{(k)}}}{\pi k(n - k)^k}.$$

The convergence is even faster, has a geometric rate, for analytic functions; we shall omit the somewhat complicated details of the constructive formulation of this theorem.

Interpolants through Chebyshev points are also nearly as good approximants as any other polynomial approximant of the same degree can be:

**Proposition 2** ([37, Theorem 16.1]). *Let  $f$  be a continuous function on  $[-1, 1]$ , and  $p_n$  its degree  $n$  interpolant through the Chebyshev points (1). Then for every polynomial  $q$  of degree  $n$ , we have*

$$\|f - p_n\| \leq \left(2 + \frac{2}{\pi} \log(n + 1)\right) \|f - q\|.$$

Specifically, as  $(2 + \frac{2}{\pi} \log(n + 1)) < 10$  for all  $n \leq 200000$ , this proposition says that in every application considered in this paper, interpolants on Chebyshev points lose at most one digit of accuracy compared to the best polynomial approximant of the same degree.

**2.2. Hermite interpolants and optimal one-sided approximation.** In some applications it is imperative that the computed approximate optimal solution  $x^*$  to (SILP) be feasible (and not just approximately feasible). However, because Lagrange interpolants usually oscillate around  $f$ , the polynomial approximation of  $A(t)x^*$  might also oscillate around the true function, and as a result, might be slightly infeasible to (SILP). If this is undesirable,

and  $x$  is componentwise nonnegative, then this problem can be avoided by using *one-sided approximations*, that is, polynomial approximants that are always below (or above) the true function. An elementary approach is to replace the approximant  $p_n$  of Proposition 1 by  $(p_n - \varepsilon)$ , or  $(p_n + \varepsilon)$  for upper approximations, where  $\varepsilon = 4V_{f^{(k)}}/(\pi k(n-k)^k)$ . With the help of Proposition 2 one can argue that this is also a near-best one-sided approximation in the uniform norm.

If the derivatives of  $f$  are of constant sign, we can also use the best one-sided approximation of  $f$  in the  $L_1$ -norm, which can be characterized as an (Hermite) interpolant at the zeros of certain orthogonal polynomials. These zeros depend only on the degree of the approximation, but not the approximated function  $f$  itself. The following proposition summarizes the characterization of these polynomial approximants. (See, for example, [12, Section 1.4] for a definition of the Legendre and Jacobi polynomials referred to in the next Proposition.)

**Proposition 3** ([6]). *Assume that  $f$  is a continuous function on  $[-1, 1]$  whose  $(n+1)$ -st derivative  $f^{(n+1)}$  is nonnegative on  $(-1, 1)$ .*

- (1) *If  $n = 2k - 1$  is odd, let  $t_1, \dots, t_k$  be the zeros of the Legendre polynomial of degree  $k$ . Then the degree- $n$  polynomial of best approximation of  $f$  from below in the  $L_1$  norm is the unique polynomial  $p_n$  satisfying*

$$p_n(t_\ell) = f(t_\ell) \quad \text{and} \quad p'_n(t_j) = f'(t_\ell), \quad \ell = 1, \dots, k.$$

- (2) *If  $n = 2k$  is even, let  $t_1, \dots, t_k$  be the zeros of the Jacobi polynomial  $P_k^{(0,1)}$ . Then the degree- $n$  polynomial of best approximation of  $f$  from below in the  $L_1$  norm is the unique polynomial  $p_n$  satisfying*

$$p_n(-1) = f(-1) \quad \text{and} \quad p_n(t_\ell) = f(t_\ell) \quad \text{and} \quad p'_n(t_\ell) = f'(t_\ell), \quad \ell = 1, \dots, k.$$

Analogous characterizations of best approximations *from above* are also given in [6].

### 3. SUM-OF-SQUARES INTERPOLANTS

We say that a polynomial is *sum-of-squares* if it can be written as a (finite) sum of squared polynomials. Specifically, we write  $p \in \text{SOS}_k$  if  $p$  is a polynomial (of degree at most  $2k$ ) that can be written as a sum of squares of polynomials of degree  $k$ . It is well-known [31] that a univariate polynomial  $p$  of degree  $2k$  is nonnegative on the entire real line if and only if  $p \in \text{SOS}_k$ . Similarly, a polynomial  $p$  of degree  $n$  is nonnegative over  $[-1, 1]$  if and only if it can be written as a weighted sum of squared polynomials [24], either in the form of

$$p(t) = (1+t)q(t) + (1-t)r(t), \quad q \in \text{SOS}_{k-1}, \quad r \in \text{SOS}_{k-1} \quad \text{if } n = 2k-1, \quad (3)$$

or in the form

$$p(t) = (1+t)(1-t)q(t) + s(t), \quad q \in \text{SOS}_{k-1}, \quad s \in \text{SOS}_k, \quad \text{if } n = 2k. \quad (4)$$

This, in turn, yields a semidefinite representation of the set of nonnegative polynomials of a fixed degree, using the fact that the cone of sums of squares of functions from any finite dimensional functional space is semidefinite representable [26, 30]. The precise form of this semidefinite representation depends on the bases that the squared polynomials being squared and the sum-of-squares polynomial  $p$  are represented in.

For the purposes of this paper, we need a representation that uses only the values of  $p$  and its derivatives at prescribed interpolation points, so that the polynomial approximations of the functions involved in (SILP) need not be explicitly constructed, but one may work

directly with sampled values of the original functions to be approximated. For Lagrange interpolants at the points  $t_0, \dots, t_n$ , this is equivalent to representing the squared polynomials in an arbitrary basis, while representing  $p$  in the Lagrange basis polynomials corresponding to the interpolation points  $t_0, \dots, t_n$ .

**Proposition 4** ([22]). *Let  $t_0, \dots, t_{2k} \in \mathbb{R}$  be distinct and  $f_0, \dots, f_{2k} \in \mathbb{R}$  be arbitrary. Fix an arbitrary basis  $p_0, \dots, p_k$  of polynomials of degree  $k$ . Then there is some polynomial  $q \in \text{SOS}_k$  interpolating each  $(t_\ell, f_\ell)$  if and only if there exists a  $(k+1) \times (k+1)$  positive semidefinite matrix  $X$  satisfying*

$$A^{(\ell)} \bullet X = f_\ell \quad \ell = 0, \dots, 2k, \quad \text{where} \quad A_{ij}^{(\ell)} = p_i(t_\ell)p_j(t_\ell). \quad (5)$$

We omit the proof, as this proposition is subsumed by Theorem 5 below, which provides a similar characterization for Hermite interpolants. The following is a generalization of both Proposition 4 above and (the main, unnumbered, results of) Sections 3.2 and 3.3 of [9].

**Theorem 5.** *Let  $t_1, \dots, t_k \in \mathbb{R}$  be distinct, let  $m_0, \dots, m_k$  and  $d$  be nonnegative integers satisfying  $2d+1 = \sum_{\ell=1}^k (m_\ell + 1)$ , and let  $f_\ell^{(m)} \in \mathbb{R}$  be arbitrary for every  $\ell = 1, \dots, k$  and  $m = 0, \dots, m_\ell$ . Also fix an arbitrary basis  $p_0, \dots, p_d$  of polynomials of degree  $d$ . Then there is some polynomial  $q \in \text{SOS}_d$  satisfying*

$$q^{(m)}(t_\ell) = f_\ell^{(m)} \quad \ell = 1, \dots, k, \quad m = 0, \dots, m_\ell$$

*if and only if there exists a positive semidefinite matrix  $X$  satisfying*

$$A^{(\ell, m)} \bullet X = f_\ell^{(m)} \quad \ell = 0, \dots, 2k, \quad \text{where} \quad A_{ij}^{(\ell, m)} = \frac{d^m}{dt^m} (p_i(t)p_j(t)) \Big|_{t=t_\ell}. \quad (6)$$

*Proof.* Using the shorthand  $p(t)$  to denote the column vector  $(p_0(t), \dots, p_d(t))^T$ ,  $q \in \text{SOS}_d$  if and only if there exists some  $d \times d$  positive semidefinite matrix  $X$  with which  $q(t) = p(t)^T X p(t) = (p(t)p(t)^T) \bullet X$  for every  $t \in \mathbb{R}$ . Differentiating both sides of this equation, we obtain

$$\frac{d^m}{dt^m} q(t) = \frac{d^m}{dt^m} (p(t)p(t)^T) \bullet X \quad \forall t, \quad (7)$$

where the differentiation of the matrix on the right-hand side is understood componentwise.

Since the prescribed derivative values and the degree determine  $q$  uniquely, equation (7) holds for every  $m \geq 0$  and every  $t \in \mathbb{R}$  if and only if it holds for each  $t_\ell$ ,  $\ell = 0, \dots, k$  with  $m$  from 0 up to  $m_\ell$ , which is precisely equation (6) in our claim.  $\square$

#### 4. SCALING

One of the key difficulties in working with polynomials of high degree is that numerical difficulties may arise if the polynomials involved are represented in an unsuitable basis, such as the monomial basis, as it is customary in the sum-of-squares literature. In the representation of Proposition 4 one can freely choose both the interpolation points and the basis  $p$ . The choice of Chebyshev points of the first kind and appropriately scaled Chebyshev polynomials works particularly well. Recall that the *Chebyshev polynomials of the first kind* are the sequence of polynomials of increasing degree defined by the recursion

$$T_0(t) = 1, \quad T_1(t) = t, \quad T_i(t) = 2tT_{i-1}(t) - T_{i-2}(t) \quad i = 2, 3, \dots \quad (8)$$

The following lemma states that if we represent the polynomials to be squared in the appropriately scaled Chebyshev basis, and use Chebyshev points as the interpolation points, then the representation (5) is perfectly scaled.

**Lemma 6.** *Let  $t_0, \dots, t_{2k}$  be the Chebyshev points given in (2) (with  $n = 2k$ ), and define  $p_0 = \sqrt{\frac{1}{2k+1}}T_0$  and  $p_i = \sqrt{\frac{2}{2k+1}}T_i$  for  $i = 1, \dots, k$ , where the  $T_i$  are the Chebyshev polynomials given in (8). Then the vectors  $(p_i(t_0), \dots, p_i(t_{2k}))$  for  $i = 0, \dots, k$  form an orthonormal system.*

*Proof.* The statement is an easy consequence of the well-known *discrete orthogonality relation* of Chebyshev polynomials [15, eq. (3.30)],

$$\sum_{\ell=0}^n T_i(t_\ell)T_j(t_\ell) = K_i\delta_{ij},$$

where  $K_0 = n + 1$ ,  $K_i = (n + 1)/2$  when  $i \geq 1$ , and  $\delta$  is the Kronecker symbol. Applying this identity, we have the following:

(1) If  $i = j = 0$ , then

$$\sum_{\ell=0}^{2k} p_0^2(t_\ell) = \frac{1}{2k+1} \sum_{\ell=0}^{2k} T_0^2(t_\ell) = 1.$$

(2) If  $i > j = 0$ , then

$$\sum_{\ell=0}^{2k} p_i(t_\ell)p_j(t_\ell) = \frac{\sqrt{2}}{2k+1} \sum_{\ell=0}^{2k} T_i(t_\ell)T_0(t_\ell) = 0.$$

(3) If  $i \geq 1$  and  $j \geq 1$ , then

$$\sum_{\ell=0}^{2k} p_i(t_\ell)p_j(t_\ell) = \frac{2}{2k+1} \sum_{\ell=0}^{2k} T_i(t_\ell)T_j(t_\ell) = \delta_{ij}. \quad \square$$

We can also express this relation in terms of the dual constraints. If  $y_\ell$  denotes the dual variable corresponding to the linear equation  $A^{(\ell)} \bullet X = f_\ell$  in (5), then dual constraint corresponding to the primal variable  $X$  is that the matrix

$$Y(y) \stackrel{\text{def}}{=} \sum_{\ell=0}^{2k} y_\ell p(t_\ell)p(t_\ell)^\top = P^\top \text{diag}(y)P \quad (9)$$

is positive semidefinite, where  $P = (p_i(t_\ell))_{\ell,i}$  is a matrix with orthonormal columns according to Lemma 6. Factoring, or computing eigenvalues of  $Y(y)$  for different values of  $y$  is therefore not numerically challenging even if the number of interpolation points (and the degree of the polynomials involved) is in the thousands. (See Section 6 for numerical examples.) An additional advantage is that algorithms that compute the values of Chebyshev polynomials at Chebyshev points to arbitrary accuracy are readily available [8, 37, 11]; also note that these values  $T_i(t_\ell)$ , and therefore the coefficient matrices  $A^{(\ell)}$  in (5) need only be computed once, offline, for every value of  $n$ .

The case of general interpolation points is in principle similarly easy [22]. For every set of points  $t_0, \dots, t_{2k}$  one can find a basis  $p_0, \dots, p_k$  of polynomials of degree  $k$  satisfying the discrete orthogonality relation

$$\sum_{\ell=0}^n p_i(t_\ell)p_j(t_\ell) = \delta_{ij},$$

by taking an arbitrary basis, and applying an orthogonalization procedure, e.g., QR factorization [17, Sec. 19]. It is important to note that the representation (5) does *not* require that the basis  $p_0, \dots, p_k$  is explicitly identified or expressed in any particular basis, only



the values of the basis polynomials are needed at the interpolation points. Throughout the orthogonalization procedure one can work directly with the values of the basis polynomials at the prescribed points. For example, the initial basis can be the Chebyshev polynomial basis, as in that basis stable evaluation of the basis polynomials is easy [8, 11], and then the orthogonalization procedure applied to the vectors of function values directly computes the values of  $p_i(t_\ell)$  for each interpolation point  $t_\ell$  for the orthogonalized basis  $p$ .

The same procedure is applicable to the weighted-sum-of-squares representations (3) and (4) of polynomials that are nonnegative over an interval. In the dual constraint (9), the  $(\ell, i)$ -th entry  $p_i(t_\ell)$  of the coefficient matrix  $P$  is replaced by  $w(t_\ell)^{1/2}p_i(t_\ell)$ , where  $w$  is the weight polynomial (in the case of polynomials over  $[-1, 1]$ , this is the polynomial  $1 - t$ ,  $1 + t$ , or  $1 - t^2$ , depending on the parity of the degree), and it is this weighted coefficient matrix that needs to be orthogonalized for a perfectly scaled representation of the weighted-sum-of-squares constraint.

## 5. HIGH-DEGREE REPRESENTATION OF LOW-DEGREE POLYNOMIALS

If the polynomial approximations of (SILP) involve interpolants of different degrees, it may be necessary to lift the lower degree interpolants into the space of higher degree ones. When polynomials are represented in the monomial basis or in an orthogonal basis, this is straightforward: the coefficients of the higher degree terms simply need to be set to zero. The analogous constraint for interpolants is that the low-degree polynomial must take consistent values at the high-degree interpolants' interpolation points, as determined by the values at the low-degree interpolation points. Since the evaluation of a polynomial at a given point is a linear functional, the mapping  $B: \mathbb{R}^{n+1} \mapsto \mathbb{R}^{N+1}$  from the values of a degree- $n$  polynomial on a given set of  $n + 1$  interpolation points to the values on a larger set of  $N + 1$  interpolation points is linear, therefore this mapping can be represented by a system of linear equality constraints.

For example, suppose that the infinite constraint set can be written in the form

$$p(t) - q(t) \geq 0 \quad \forall t \in [-1, 1],$$

where  $q$  is a given degree- $N$  polynomial represented as an interpolant on the  $N + 1$  Chebyshev points of the second kind (1), whereas  $p$  is an degree- $n$  polynomial to be optimized, with  $n < N$ , represented as an interpolant on  $n + 1$  Chebyshev points. To represent this constraint using sum-of-squares interpolants,  $p - q$  needs to be a degree- $N$  sum-of-squares interpolant, and the variable  $p$  must be “upsampled” and represented as a degree- $N$  interpolant.

The coefficient matrix of these constraints can be determined using interpolation formulae. It is important to note that although the previous section shows that the sum-of-squares representation of nonnegative interpolants can always be scaled, regardless of the location of the interpolation points, the problem of polynomial interpolation is inherently ill-conditioned in general, meaning that small changes in the values of the degree- $n$  polynomial can result in large changes in the upsampled values [4]. On the other hand, if the low-degree polynomial is an interpolant on Chebyshev points, or any other point set with asymptotic density  $(1 - x^2)^{-1/2}$ , then the interpolation problem is well-conditioned; moreover, the coefficients of the upsampling constraints can be computed efficiently and in a numerically stable manner, using barycentric Lagrange interpolation [18].

It is for this reason that in all our numerical examples in this paper, all polynomials are represented as interpolants using Chebyshev points as interpolation points.



## 6. APPLICATIONS AND NUMERICAL EXPERIMENTS

**6.1. Polynomial envelopes.** Our first example demonstrates that the computational infrastructure presented in Sections 3 and 4 is indeed capable of handling high-degree polynomials without any numerical difficulties. Consider the following problem: given degree- $d$  polynomials  $p_1, \dots, p_m$ , find the greatest degree- $n$  polynomial lower approximation of  $\min(p_1, \dots, p_m)$ , where the minimum is understood pointwise. Formally, we seek the optimal solution to

$$\begin{aligned} & \text{maximize}_p \quad \int_{-1}^1 p(t) dt \\ & \text{subject to} \quad p(t) \leq p_i(t) \quad \forall t \in [-1, 1] \quad i = 1, \dots, m. \end{aligned} \quad (10)$$

All polynomials involved can be represented as interpolants on the same  $\max(n, d) + 1$  points. The decision variables are the function values  $p(t_\ell)$ ,  $\ell = 1, \dots, \max(n, d) + 1$  at the interpolation points  $t_\ell$ . The nonnegativity of the polynomials  $p_i(t) - p(t)$  can be formulated as these polynomials being weighted sums of squares. For this example, we assume that  $n \geq d$  so that we do not have to worry about the upsampling issue discussed in the previous section.

Random instances were generated by drawing uniformly random integer coefficients from  $[-9, 9]$  for each  $p_i$  represented in the Chebyshev basis. For efficiency and scaling, the dual problem,

$$\begin{aligned} & \text{minimize}_y \quad \sum_{i, \ell} p_i(t_\ell) y_{i\ell} \\ & \text{subject to} \quad \sum_i y_{i\ell} = w_\ell \quad \forall \ell, \\ & \quad \sum_\ell (1 + t_\ell) A^{(\ell)} y_{i\ell} \geq 0 \quad \forall i, \\ & \quad \sum_\ell (1 - t_\ell) A^{(\ell)} y_{i\ell} \geq 0 \quad \forall i \end{aligned} \quad (11)$$

was solved, after orthogonalizing the matrix inequalities using the procedure outlined in Section 4. We employed three different solvers, Sedumi [34], SDPT3 version 4 [36], and CSDP version 6.2 [7], each running in Matlab 2014a, to confirm that the semidefinite formulations can indeed be solved with off-the-shelf SDP solvers, for a variety of values for the number of polynomials  $m$  as well as the degrees  $n$  and  $d$ .

In each case, the optimal polynomial interpolant  $p$  was recovered from the optimal dual solution:  $p(t_\ell)$  is the optimal dual variable corresponding to the linear equality constraint  $\sum_i y_{i\ell} = w_\ell$  in (11).

For solvers that do not handle linear equality constraints well (specifically, those that represent  $Ax = b$  as a pair of inequalities  $b \leq Ax \leq b$ ), it is useful to note that the polynomials  $p_i$  can be assumed to satisfy  $p_i(t_\ell) \leq 0$  for all  $i$  and  $\ell$ , without any loss of generality, which in turn allows us to assume that the primal variables  $p(t_\ell)$  in (10) are constrained (redundantly) to be non-positive. With this modification, the equality constraint in the dual problem (11) becomes an inequality  $\sum_i y_{i\ell} \leq w_\ell \quad \forall \ell$ .

**Example 1.** Figure 1 depicts three quintic polynomials, along with three best polynomial lower approximations of their pointwise minimum, of increasing degrees (degree 5, 15, and 75). The 75-degree lower approximation is visually nearly indistinguishable from the

minimum of the three polynomials. The computations for this plot were carried out using Sedumi.

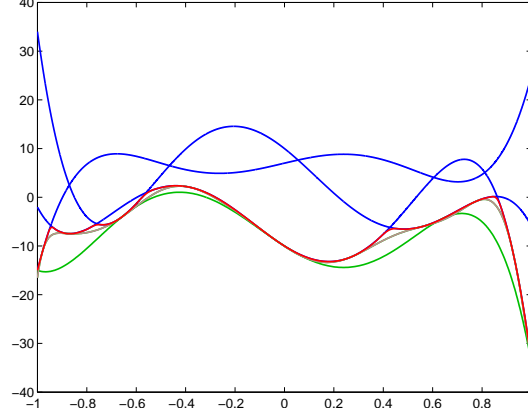


FIGURE 1. Three polynomials of degree 5, along with their best lower polynomial approximations of degree 5, 15, and 75.

To find the optimal polynomials with the highest numerically possible accuracy, we set the Sedumi accuracy goal `eps` to zero so that the solver keeps iterating as long as it can make any progress. Figure 2 shows the plot of the difference between  $\min_i p_i$  and the three polynomial lower approximations. Only the sections of the plots close to the  $x$ -axis are shown, in order to demonstrate that the resulting optimal polynomials are computed to sufficiently high accuracy that the points of contact (the points where  $\min_i p_i(t) = p(t)$ ) can be separated, and computed to several digits of precision.

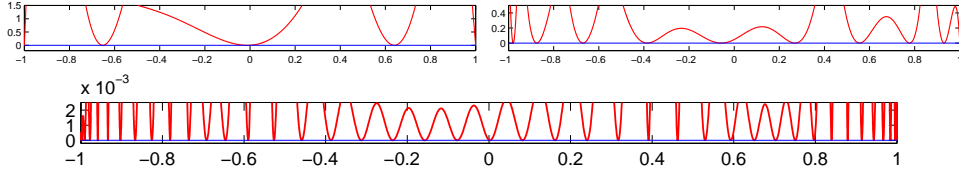


FIGURE 2. Pointwise difference between the three optimal polynomials and  $\min_i p_i$  from Figure 1. Only the near-zero section of the plots are shown; the number of contact points can be easily read off the diagrams. The polynomials have degrees 5, 15, and 75, respectively.

To test the limits of the approach when applied to polynomials of very high degree, similar problems were solved for higher values of  $n$ , with the three SDP solvers mentioned above (Sedumi, SDPT3, and CSDP). As before, to obtain the highest possible accuracy, we set tolerances and accuracy goals to zero so that the solvers keep iterating as long as they can make any progress. Otherwise, default parameter settings were used with both solvers.

With each solver, as the sizes of the SDPs grow quadratically with the degree of the polynomials involved, the available memory became a bottleneck. Therefore, we reduced the number of constraints to  $m = 2$ , and then increased  $n$  as shown in the tables below. Using a standard desktop computer with 32GB RAM, the degree was increased until the

$n + 1$	# of nonzeros	# of iterations	solver time [s]	primal inf.	dual inf.	duality gap
100	0.5 M	23	5	$5.0 \cdot 10^{-10}$	$3.0 \cdot 10^{-14}$	$3.25 \cdot 10^{-14}$
200	4.0 M	21	44	$3.5 \cdot 10^{-10}$	$1.5 \cdot 10^{-13}$	$9.04 \cdot 10^{-13}$
300	13.5 M	24	215	$1.7 \cdot 10^{-10}$	$1.4 \cdot 10^{-14}$	$6.23 \cdot 10^{-15}$
400	32.1 M	21	547	$2.3 \cdot 10^{-10}$	$1.7 \cdot 10^{-13}$	$6.01 \cdot 10^{-14}$
500	62.6 M	19	1128	$1.1 \cdot 10^{-9}$	$9.1 \cdot 10^{-13}$	$2.43 \cdot 10^{-13}$
600	108 M	20	2456	$2.6 \cdot 10^{-9}$	$2.0 \cdot 10^{-12}$	$4.56 \cdot 10^{-13}$
700	171 M	21	4847	$4.8 \cdot 10^{-10}$	$3.0 \cdot 10^{-13}$	$6.19 \cdot 10^{-14}$
800	256 M	21	8670	$7.2 \cdot 10^{-10}$	$3.2 \cdot 10^{-13}$	$5.76 \cdot 10^{-14}$
900	321 M	20	12969	$1.9 \cdot 10^{-9}$	$1.1 \cdot 10^{-12}$	$1.80 \cdot 10^{-13}$
1000	501 M	19	19875	$5 \cdot 10^{-9}$	$2.8 \cdot 10^{-12}$	$4.19 \cdot 10^{-13}$

TABLE 1. Solver statistics from Sedumi from the solution of Example 1. Instances of the optimization problem (10)-(11) was solved for  $m = 2$ ,  $d = 5$ , and different values of the degree  $n$ . (That is,  $n + 1$  in the heading is the number of interpolation points.) M in the second column stands for millions. The last three columns show the relative and infeasibility of the optimal primal and dual solutions, and the relative duality gap. Larger problems ( $n + 1 \geq 1100$ ) could not be solved because of memory constraints.

$n + 1$	# of iterations	solver time [s]	primal inf.	dual inf.	duality gap
100	24	4	$6.3 \cdot 10^{-10}$	$1.1 \cdot 10^{-11}$	$5.2 \cdot 10^{-12}$
200	25	25	$1.4 \cdot 10^{-9}$	$3.9 \cdot 10^{-12}$	$5.7 \cdot 10^{-12}$
300	29	107	$8.5 \cdot 10^{-9}$	$1.0 \cdot 10^{-12}$	$1.5 \cdot 10^{-11}$
400	26	264	$2.7 \cdot 10^{-9}$	$5.1 \cdot 10^{-12}$	$1.7 \cdot 10^{-11}$
500	29	695	$3.4 \cdot 10^{-9}$	$4.3 \cdot 10^{-13}$	$1.6 \cdot 10^{-11}$
600	30	1395	$9.7 \cdot 10^{-10}$	$1.6 \cdot 10^{-12}$	$3.1 \cdot 10^{-10}$
700	30	2527	$2.2 \cdot 10^{-9}$	$9.5 \cdot 10^{-13}$	$1.7 \cdot 10^{-10}$
800	33	4732	$3.0 \cdot 10^{-8}$	$2.3 \cdot 10^{-13}$	$5.4 \cdot 10^{-12}$
900	30	6724	$5.6 \cdot 10^{-10}$	$4.2 \cdot 10^{-12}$	$9.1 \cdot 10^{-10}$
1000	31	10505	$3.9 \cdot 10^{-10}$	$2.2 \cdot 10^{-13}$	$2.4 \cdot 10^{-11}$

TABLE 2. Solver statistics from SDPT3 from the solution of Example 1. Instances of the optimization problem (10)-(11) was solved for  $m = 2$ ,  $d = 5$ , and different values of the degree  $n$ . Larger problems ( $n + 1 \geq 1100$ ) could not be solved because of memory constraints.

solvers ran out of memory, and were unable to solve the problem. The number of nonzeros in the constraint matrix of the semidefinite program, along with the number of iterations, the solver running time, and the final duality gap for each run of Sedumi is shown in Table 1; the same solver statistics (without repeating the problem statistics) for SDPT3 are shown in Table 2, and in Table 3 for CSDP. It is apparent from the results that the solvers are able to solve even the largest instances, involving polynomials of degree 1000, without any numerical difficulty, and the memory constraint is the only bottleneck.

$n + 1$	# of iterations	solver time [s]	primal inf.	dual inf.	duality gap
100	17	1	$1.89 \cdot 10^{-11}$	$6.68 \cdot 10^{-13}$	$1.74 \cdot 10^{-9}$
200	19	10	$2.57 \cdot 10^{-12}$	$1.63 \cdot 10^{-12}$	$3.71 \cdot 10^{-10}$
300	21	45	$9.13 \cdot 10^{-12}$	$2.29 \cdot 10^{-10}$	$1.70 \cdot 10^{-9}$
400	19	136	$1.83 \cdot 10^{-11}$	$7.02 \cdot 10^{-12}$	$6.38 \cdot 10^{-9}$
500	21	371	$4.41 \cdot 10^{-12}$	$7.83 \cdot 10^{-10}$	$1.72 \cdot 10^{-9}$
600	23	788	$4.79 \cdot 10^{-12}$	$1.23 \cdot 10^{-10}$	$1.59 \cdot 10^{-9}$
700	22	1486	$1.54 \cdot 10^{-12}$	$3.09 \cdot 10^{-10}$	$8.39 \cdot 10^{-10}$
800	22	2474	$3.33 \cdot 10^{-12}$	$1.72 \cdot 10^{-9}$	$1.97 \cdot 10^{-9}$
900	20	4569	$7.25 \cdot 10^{-12}$	$3.57 \cdot 10^{-11}$	$7.38 \cdot 10^{-9}$
1000	22	8634	$9.00 \cdot 10^{-13}$	$1.84 \cdot 10^{-10}$	$5.68 \cdot 10^{-10}$
1100	22	15129	$9.88 \cdot 10^{-13}$	$4.46 \cdot 10^{-9}$	$7.75 \cdot 10^{-10}$

TABLE 3. Solver statistics from CSDP from the solution of Example 1. Instances of the optimization problem (10)-(11) was solved for  $m = 2$ ,  $d = 5$ , and different values of the degree  $n$ . Larger problems ( $n + 1 \geq 1200$ ) could not be solved because of memory constraints.

**6.2. Recovering best one-sided approximations.** The following example combines the semidefinite optimization model of Example 1 with high-degree polynomial approximation of nonpolynomial functions, and it also incorporates the ideas introduced in Section 5. For easier verification of the results, we pose a problem that can be solved in essentially closed form using Proposition 3.

**Example 2.** Consider the problem of finding the best polynomial lower approximation (of a given degree) of the function  $f(t) = \exp(t^{100})$  defined over  $[-1, 1]$ . It can be shown using Propositions 1 and 2 that the polynomial interpolant  $p_{200}$  of  $f$  on the 200 Chebyshev points has a maximum absolute error smaller than double machine precision. For numerical purposes, finding the best polynomial lower approximant  $p$  (of a given degree lower than 200) to  $f$  is equivalent to finding the optimal solution  $p$  to the problem

$$\begin{aligned}
& \text{maximize}_p \quad \int_{-1}^1 p(t) dt \\
& \text{subject to} \quad p(t) \leq p_{200}(t) \quad \forall t \in [-1, 1].
\end{aligned} \tag{12}$$

This problem can be translated to a semidefinite program nearly identically to the previous example, except that now  $p$  has to be upsampled to the 200 interpolation points, using

barycentric interpolation, as discussed in Section 5. The modified dual problem becomes

$$\begin{aligned}
& \text{minimize}_{y,z} && \sum_{i,\ell} p_i(t_\ell) y_{i\ell} \\
& \text{subject to} && \sum_i y_{i\ell} + z_\ell = w_\ell \quad \forall \ell, \\
& && \sum_\ell (1 + t_\ell) A^{(\ell)} y_{i\ell} \succcurlyeq 0 \quad \forall i, \\
& && \sum_\ell (1 - t_\ell) A^{(\ell)} y_{i\ell} \succcurlyeq 0 \quad \forall i, \\
& && B^T z = 0,
\end{aligned} \tag{13}$$

where  $B \in \mathbb{R}^{200 \times n}$  is the barycentric interpolation matrix mapping the values of a degree- $(n-1)$  polynomial  $p$  at the  $n$  Chebyshev points to the values at the 200 Chebyshev points used to represent  $p_{200}$ .

As an example, we solved this problem to determine the optimal polynomial lower approximant of degree 49 (represented as an interpolant on the 50 Chebyshev points) using SeDuMi. As before, for the highest numerically possible accuracy, we set the SeDuMi accuracy goal `eps` to zero so that the solver iterates while it can make any progress. Finally, the points of contact of the roots of optimal approximant were determined numerically using the root finding algorithm for interpolants implemented in the Matlab toolbox Chebfun v.5.0.1 [11]. The obtained roots are shown in Table 4.

Proposition 3 provides a characterization of the optimal solution as an Hermite interpolant on the roots of the 25-degree Legendre polynomial  $L_{25}$ , meaning the points of contact are the (known, and numerically precisely computable) roots of  $L_{25}$ , allowing us to check the accuracy of our calculations. Table 4 shows the values of the roots obtained from our optimization procedure and the correct values with machine precision accuracy. The largest absolute error of the roots was  $6.16 \cdot 10^{-7}$ , the largest relative error (not defined for the root equal to zero) was  $4.88 \cdot 10^{-6}$ . In other words, all roots were accurate up to at least 5 significant digits.

**6.3. Experimental design.** Problems discussed in this section are among the author’s main motivation for studying semi-infinite linear programs with additional conic constraints.

The goal of optimal experimental design in general is to maximize the quality of statistical inference by collecting the right data, given limited resources. In the context of linear regression, the inference is based on a data model

$$y(t) = \sum_{i=1}^m \beta_i f_i(t) + \epsilon(t), \tag{14}$$

where  $f_1, \dots, f_m$  are known functions, and the random variable  $\epsilon$  (of known probability distribution) represents measurement errors and other sources of variation unexplained by the model. In the experiment, the (noisy) values of  $y$  are observed for a number of different values of  $t$  chosen from the given *design space*  $\mathcal{I}$ , and the goal of the experiment is to infer the values of the unknown coefficients  $\beta_i$ ,  $i = 1, \dots, m$ . By a *design* we mean a set of values  $\{t_1, \dots, t_s\}$  for which the response  $y(t_i)$  is to be measured, along with the number of repeated measurements  $r_i$  to be taken at each  $t_i$ . Multiple measurements are allowed, since assuming independent measurement errors, the repeated measurements can help reduce our uncertainty of the “true” (noise-free) value of each  $y(t_i)$ . The problem of deciding how

Computed point of contact	Exact point of contact
-0.995556972963306	-0.995556969790498
-0.976663935477085	-0.976663921459518
-0.942974611506432	-0.942974571228974
-0.894992079192159	-0.894991997878275
-0.833442768114398	-0.833442628760834
-0.75925946688898	-0.759259263037358
-0.673566645603713	-0.673566368473468
-0.57766328506073	-0.577662930241223
-0.473003173358265	-0.473002731445715
-0.361172781372549	-0.361172305809388
-0.243867464225739	-0.243866883720988
-0.122865277764643	-0.12286469261071
-6.15973190950935 $\cdot 10^{-7}$	0
0.122864093106243	0.12286469261071
0.243866329210549	0.243866883720988
0.361171807947986	0.361172305809388
0.473002281718963	0.473002731445715
0.577662568403695	0.577662930241223
0.673566077288833	0.673566368473468
0.759259051474652	0.759259263037358
0.833442487648842	0.833442628760834
0.894991911861248	0.894991997878275
0.942974528231816	0.942974571228974
0.976663905379173	0.976663921459518
0.995556966216301	0.995556969790498

TABLE 4. Comparison of the numerically computed points of contact from Example 2 and the exact values (shown with double machine precision accuracy) derived from Proposition 3. In spite of the high degree of the polynomials involved, all computed points of contact (computed as the roots of high-degree sum-of-squares interpolants) are accurate up to at least 5 significant digits.

many (discrete) measurements to take at what points  $t_i$  is a non-convex (combinatorial) optimization problem, which is commonly simplified to a convex problem by relaxing the integrality constraints on  $r_i$  [14, 32]. In the resulting model one can normalize the vector  $r$  by assuming  $\sum_i r_i = 1$  (in addition to  $r \geq 0$ ), so that  $r_i$  represents not the number, but the fraction of experiments to be conducted at point  $t_i$ .

This way, the experiment design is mathematically a finitely supported probability distribution  $\xi$  satisfying  $\xi = t_i$  with probability  $r_i$ ,  $i = 1, \dots, s$ . It is immediate that the feasible set (the set of probability measures supported on a finite subset of a given set  $\mathcal{I} \subseteq \mathbb{R}^n$ ) is convex.

Our goal with the experiment is to maximize our confidence in the estimated components of  $\beta$ , which can be quantified using the Fisher information [14, 32] that the measured values carry about  $\beta$ . Under the common assumption that  $\epsilon(t)$  is normally distributed with mean

zero and variance  $\omega(t)$ , we can estimate the parameters  $\beta$  using ordinary least-squares. Using the notation  $f(t) = (f_1(t), \dots, f_m(t))^T$ , the *Fisher information matrix* of  $\beta$  corresponding to the design  $\xi$  is

$$M(\xi) = \int_{\mathcal{I}} f(t)f(t)^T \omega(t) d\xi(t). \quad (15)$$

Of course, this integral simplifies to a finite sum for every design. Note that for every  $\xi$ ,  $M(\xi) \in \mathbb{S}_+^m$ , therefore the optimization usually takes place with respect to some *optimality criterion*  $\Phi$  that measures the quality of the Fisher information matrix. If  $\Phi$  is an  $\mathbb{S}_+^m \mapsto \mathbb{R}$  function, the design  $\hat{\xi}$  is called *optimal with respect to  $\Phi$* , or  $\Phi$ -*optimal* for short, if  $\Phi(M(\hat{\xi}))$  is maximum. Only those criteria  $\Phi$  are interesting that are compatible with the Löwner partial order, that is functions  $\Phi$  satisfying  $\Phi(A) \geq \Phi(B)$  whenever  $A \succcurlyeq B \succcurlyeq 0$ ; see [32, Chap. 4] for a statistical interpretation of this requirement. Popular choices of  $\Phi$  include  $\Phi(M) = \det(M)$ ,  $\Phi(M) = \lambda_1(M)$  (smallest eigenvalue),  $\Phi(M) = -\text{tr}(M^{-1})$ , and  $\Phi(M) = (\text{tr}(M^p))^{1/p}$  for  $p \geq 1$ .

The main result (Theorem 2) of [28] is that if the basis functions  $f_i$  in (14) are rational functions, then the optimal design can be computed by solving two semidefinite programs. (For completeness, the result is repeated in the Appendix.) The first semidefinite program determines a polynomial whose roots are the support points of the optimal design, while the second one is used to determine the probability masses assigned to the support points once the support points are known. The first semidefinite program involves a constraint that a polynomial be nonnegative over the design space  $\mathcal{I}$ ; see Theorem 7 in the Appendix. Therefore, this is an instance of (SILP) that cannot be handled effectively using the commonly used cutting plane methods of semi-infinite linear programming.

Most practical problems involve basis functions that are not polynomials or rational functions, therefore we follow the approach suggested in Section 1, and replace each  $f_i$  by a close polynomial (or rational) approximant. If any of these approximants has a high degree, then the aforementioned semidefinite program determines a high-degree polynomial that must be nonnegative over  $\mathcal{I}$ , and that must be computed with sufficient accuracy that allows its roots to be precisely computed.

**Example 3.** Consider the linear regression model (14) involving a mixture of  $m = 3$  Gaussians  $f_i = \exp(-3(x - \mu_i)^2)$  with  $\mu_1 = -0.5$ ,  $\mu_2 = 0$ , and  $\mu_3 = 0.5$ , and suppose we are interested in finding the best E-optimal design for determining the best fit, over the design space  $\mathcal{I} = [-1, 1]$ . Since the  $f_i$  are not polynomials, we will approximate them by high-degree polynomial interpolants.

E-optimality means maximizing the smallest eigenvalue  $\Phi = \lambda_1$  of the Fisher information matrix. Therefore, this is a semidefinite representable optimality criterion:

$$\lambda_1(M) \geq z \quad \Longleftrightarrow \quad M - Iz \succcurlyeq 0.$$

Invoking Theorem 7 in the Appendix with  $p = 1$ ,  $A_1 = \text{id}$ ,  $B_1 = I$ ,  $C_1(\cdot) = 0$ , and  $D_1 = 0$  in (16), we obtain that the support of the optimal design is a subset of the roots of the optimal polynomial  $\pi$  determined by the solution of the optimization problem

$$\begin{aligned} & \underset{y \in \mathbb{R}, \pi \in \mathbb{R}^{d+1}, W \in \mathbb{S}_+^3}{\text{minimize}} && y \\ & \text{subject to} && \text{tr}(W) = -1 \\ & && \pi(t) \stackrel{\text{def}}{=} y - W \bullet M_t \geq 0 \quad \forall t \in [-1, 1], \end{aligned}$$

where  $M_t = f(t)f(t)^T$  with  $f(t) = (f_1(t), f_2(t), f_3(t))^T$ .



Using Chebfun (or the bound of Proposition 1), we obtain that all nonpolynomial functions involved in the optimization problem (including not only  $f_i$ , but the products  $f_i f_j$ ) can be approximated within machine-precision uniform error over  $[-1, 1]$  by polynomial interpolants on 40 Chebyshev points. The nonnegativity constraint is replaced by the constraint that  $\pi$  is weighted-sum-of-squares with weights  $1 + t$  and  $1 - t$ .

As in the previous examples, we solved the resulting semidefinite program, and obtained the optimal polynomial shown on Figure 3. The polynomial has three roots in  $[-1, 1]$ , these are  $\pm 0.7410$  and 0.

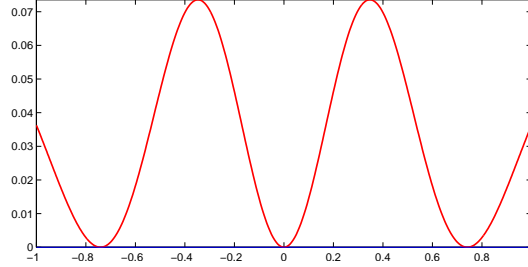


FIGURE 3. The optimal polynomial  $\pi$  of degree 40 from the optimal design of experiments problem discussed in Example 3. The optimal design is supported on the roots of this polynomial located in  $[-1, 1]$ , which are approximately  $\pm 0.7410$  and 0.

This example was also implemented in Matlab, and solved with multiple solvers. Neither solver reported any errors or warnings during the solution, and returned the same solution (within the expected accuracy).

6.3.1. *Extension to nonlinear regression models.* A similar approach can be used to compute optimal designs of experiments for nonlinear regression models

$$y(t) = f(t; \beta_1, \dots, \beta_m) + \epsilon(t).$$

For such models, there are several non-equivalent definitions of optimal designs, here we only consider the simplest ones called *local optimal designs*. (These are not local optimal solutions to a nonconvex optimization model; the term “local” has a statistical meaning in this context that will be clarified below.)

Local optimal designs for nonlinear models are defined in an almost identical manner to optimal designs for linear models, except that the in the definition (15) of the Fisher information matrix, the matrix  $f(t)f(t)^T$  is replaced by the matrix

$$(\partial f(t)/\partial \beta)(\partial f(t)/\partial \beta)^T,$$

where  $\partial f(t)/\partial \beta$  denotes the vector of partial derivatives  $\partial f(t)/\partial \beta_i$ ,  $i = 1, \dots, m$ . It can be seen that for linear models,  $\partial f(t)/\partial \beta$  is simply  $f$ , and we arrive at the previous definition. However, for nonlinear models,  $\partial f(t)/\partial \beta$  is dependent on the unknown parameters  $\beta$  that the experiment is designed to identify. Therefore, local optimal designs are defined with respect to an initial guess  $\hat{\beta}$  of these parameter values; replacing the true Fisher information matrix with a local estimate,

$$M_{\hat{\beta}}(\xi) = \int_{\mathcal{I}} \left( \frac{\partial f(t)}{\partial \beta} \right) \left( \frac{\partial f(t)}{\partial \beta} \right)^T \Big|_{\beta=\hat{\beta}} \omega(t) d\xi(t).$$

A design  $\hat{\xi}$  is said to be *locally optimal with respect to  $\Phi$* , or *locally  $\Phi$ -optimal* for short, if  $\Phi(M_{\hat{\beta}}(\hat{\xi}))$  is maximum. The semidefinite programming characterization of the support of local optimal designs is entirely analogous to the characterization for linear models given in Theorem 7.

**Example 4.** We consider what is perhaps the simplest nonlinear regression model: logistic regression with only two parameters. In this model we have

$$f(t) = (1 + \exp(-\beta_0 - \beta_1 t))^{-1},$$

and the partial derivatives of  $f$  with respect to the parameters are

$$\partial f(t)/\partial \beta_0 = (2 + 2 \cosh(\beta_0 + \beta_1 t))^{-1} \quad \text{and} \quad \partial f(t)/\partial \beta_1 = t(2 + 2 \cosh(\beta_0 + \beta_1 t))^{-1}.$$

Using the shorthand  $g(t) = (2 + 2 \cosh(\hat{\beta}_0 + \hat{\beta}_1 t))^{-1}$ , the optimization model characterizing the support involves as a constraint the nonnegativity of a function that is in the space

$$\text{span}\{1, g(t), tg(t), t^2 g(t)\}.$$

For our numerical example, we chose the values  $\beta_0 = 0$  and  $\beta_1 = 12$ .

In spite of its apparent simplicity, the semidefinite program that arises from a polynomial approximation of  $g$  is far from straightforward. First, it can be shown that  $g$  is a function that exhibits the *Runge phenomenon* [13], that is, the maximum (pointwise) error between  $g$  and its polynomial interpolants on equispaced points does not tend to zero as the number of interpolation points increases. (Figure 4.) Even if we use Chebyshev points, the 100-degree interpolant of  $g$  is highly oscillatory, and yields entirely wrong results, with no correct significant digits in the optimal objective function value.

On the other hand, the 200-degree interpolant of  $g$  on Chebyshev points already has a uniform error less than machine precision over  $[-1, 1]$ . Using this approximation, we obtain that the optimal polynomial has two roots:  $\pm 0.08697$ . As in the previous examples, all of the SDP solvers returned the same solution without numerical problems, in spite of the high degree of the polynomials involved in the computation.

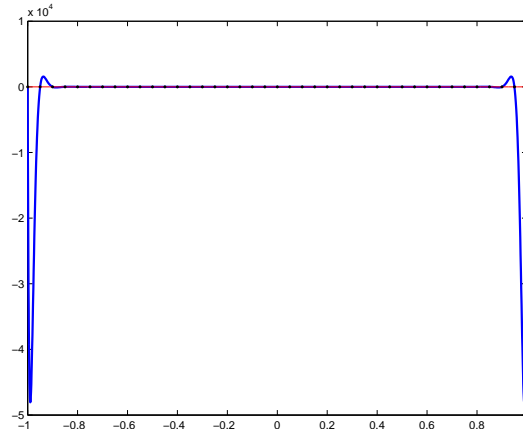


FIGURE 4. The function  $g$  from Example 4 exhibits the Runge phenomenon. Shown are the function  $g$  in red (this bell-shaped curve appears to be a horizontal line because of the scale) and its polynomial interpolant on 40 equispaced points.

## 7. DISCUSSION

While the examples presented in the previous section are admittedly toy examples of their respective application areas, the semidefinite programs they lead to are impossible to solve with commonly used semidefinite solvers if the nonnegative polynomial constraints are translated to semidefinite constraints in the standard fashion, due to the high degrees of the polynomials involved. Example 4 in particular demonstrates that the possibly simplest problem in optimal design of experiments for nonlinear statistical models (the two-parameter version of the most studied nonlinear regression model) involves nonpolynomial functions that cannot be adequately approximated with polynomials of less than hundred degrees, therefore there is a definite need to be able to reliably optimize over cones of sum-of-squares polynomials of hundreds of degrees.

The numerical results from these experiments, on the other hand, demonstrate that the semidefinite representations of sum-of-squares interpolants enable the solution of these problems with all of the existing semidefinite programming solvers tried in this study: CSDP, SDPT3, and SeDuMi, thanks to the considerably better numerical properties of the representation. The optimal polynomials were computed with high enough accuracy that even their roots could be located precisely—this is particularly important for their application in the design of experiments.

To derive the semidefinite programming models, it is essential that the high degree polynomial interpolants of given functions be computed quickly and accurately, and that the resulting interpolants can be processed as needed. For instance, the evaluation of interpolants at points other than the interpolation points has to be efficient and stable, and the roots of interpolants need to be computed efficiently. All the infrastructure required for these computations are already available in constructive approximation packages such as the Matlab toolbox Chebfun. Our results show that these tools can be combined with the existing semidefinite programming solvers to obtain a reliable tool that can solve semidefinite linear programs with general univariate functional constraints, without the need to derive tailor-made specialized algorithms.

Some questions remain open, mostly around the issue of efficiency. A general criticism of sum-of-squares techniques is that the semidefinite programming representations of nonnegative polynomials of degree  $n$  turn what is inherently an  $O(n)$ -variable optimization problem into a problem with  $O(n^2)$  variables. This limits somewhat the degree of polynomials that can be practically handled in sum-of-squares optimization. The techniques presented in this paper do not address this issue, only the problem of poor scaling.

That said, the semidefinite solvers could handle all instances of the problems that could fit in the 32 GBs of memory of the desktop computer used in the experiments; this means several constraints involving 1000-degree polynomials. For applications requiring polynomials of even higher degree, it will be necessary to find either sparser representations, or tailor-made methods that can handle the higher degree sum-of-squares constraints directly, avoiding the use of semidefinite representation of size  $O(n^2)$ . These might also be helpful in the generalization of the approach to multivariate polynomials.

## REFERENCES

- [1] Farid Alizadeh and Dávid Papp. Estimating arrival rate of nonhomogeneous Poisson processes with semidefinite programming. *Annals of Operations Research*, 208(1):291–308, 2013. <http://link.springer.com/article/10.1007/s10479-011-1020-2>.
- [2] Bernhard Beckermann. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numerische Mathematik*, 85(4):553–577, June 1997. doi:10.1007/PL00005392.

- [3] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, Philadelphia, PA, 2001.
- [4] Jean-Paul Berrut and Lloyd N. Trefethen. Barycentric Lagrange interpolation. *SIAM Review*, 46(3):501–517, 2004.
- [5] Bruno Betrò. An accelerated central cutting plane algorithm for linear semi-infinite programming. *Mathematical Programming*, 101:479–495, 2004.
- [6] R. Bojanic and R. DeVore. On polynomials of best one sided approximation. *L'Enseignement Mathématique*, 12(3):139–164, 1966.
- [7] B. Borchers. CSDP, a C library for semidefinite programming. *Optimization Methods & Software*, 11-2(1-4):613–623, 1999.
- [8] C. W. Clenshaw. A note on the summation of Chebyshev series. *Mathematics of Computation*, 9:118–120, 1955.
- [9] E. de Klerk, G. E. E. Elfadul, and D. den Hertog. Optimization of univariate functions on bounded intervals by interpolation and semidefinite programming. Technical Report 26, Tilburg University, April 2006.
- [10] Etienne De Klerk, EG Elabwabi, and Dick Den Hertog. Optimization of univariate functions on bounded intervals by interpolation and semidefinite programming. Technical Report Center Discussion Paper 2006-26, Tilburg University, 2006.
- [11] T. A. Driscoll, N. Hale, and L. N. Trefethen. *Chebfun Guide*. Pafnuty Publications, Oxford, UK, 2014.
- [12] Charles F. Dunkl and Yuan Xu. *Orthogonal Polynomials of Several Variables*, volume 81 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, UK, 2001.
- [13] James F. Epperson. On the Runge example. *The American Mathematical Monthly*, 94(4):329–341, 1987. doi:10.2307/2323093.
- [14] V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, NY, 1972.
- [15] Amparo Gil, Javier Segura, and Nico M Temme. *Numerical methods for special functions*. SIAM, Philadelphia, PA, 2007.
- [16] Miguel A. Goberna and Marco A. López. *Linear Semi-Infinite Optimization*. John Wiley & Sons, Chichester, UK, 1998.
- [17] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2002.
- [18] Nicholas J. Higham. The numerical stability of barycentric Lagrange interpolation. *IMA Journal of Numerical Analysis*, 24(4):547–556, 2004.
- [19] Samuel Karlin and William J. Studden. *Tchebycheff Systems, with Applications in Analysis and Statistics*, volume XV of *Pure and Applied Mathematics*. Interscience, New York, NY, 1966.
- [20] K. O. Kortanek and Hoon No. A central cutting plane algorithm for convex semi-infinite programming problems. *SIAM Journal on Optimization*, 3(4):901–918, November 1993.
- [21] Guo-xin Liu. A homotopy interior point method for semi-infinite programming problems. *Journal of Global Optimization*, 37(4):631–646, 2007. <http://dx.doi.org/10.1007/s10898-006-9077-1>.
- [22] Johan Lofberg and Pablo A Parrilo. From coefficients to samples: a new approach to sos optimization. In *43rd IEEE Conference on Decision and Control*, volume 3, pages 3154–3159. IEEE, 2004.
- [23] Marco López and Georg Still. Semi-infinite programming. *European Journal of Operational Research*, 180(2):491–518, 2007. <http://dx.doi.org/10.1016/j.ejor.2006.08.045>.
- [24] Franz Lukács. Verschärfung der ersten Mittelversatzes der Integralrechnung für rationale Polynome. *Mathematische Zeitschrift*, 2:229–305, 1918. doi:10.1007/BF01199412.
- [25] Sanjay Mehrotra and Dávid Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimizaton*, 24(4):1670–1697, 2014. <http://dx.doi.org/10.1137/130925013>.
- [26] Yurii Nesterov. Squared functional systems and optimization problems. In Hans Frenk, Kees Roos, Tamás Terlaky, and Shuzhong Zhang, editors, *High Performance Optimization*, volume 33 of *Applied Optimization*, pages 405–440. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
- [27] Dávid Papp. *Optimization models for shape-constrained function estimation problems involving nonnegative polynomials and their restrictions*. PhD thesis, Rutgers University, May 2011. [http://www4.ncsu.edu/~dpapp/pub/thesis\\_dpapp.pdf](http://www4.ncsu.edu/~dpapp/pub/thesis_dpapp.pdf).
- [28] Dávid Papp. Optimal designs for rational function regression. *Journal of the American Statistical Association*, 107(497):400–411, March 2012. <http://dx.doi.org/10.1080/01621459.2012.656035>.
- [29] Dávid Papp and Farid Alizadeh. Shape constrained estimation using nonnegative splines. *Journal of Computational and Graphical Statistics*, 23(1):211–231, 2014. <http://amstat.tandfonline.com/doi/full/10.1080/10618600.2012.707343>.

- [30] Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003. doi:10.1007/s10107-003-0387-5.
- [31] George Pólya and Gabor Szegő. *Problems and Theorems in Analysis*, volume 2. Springer, New York, NY, 1976.
- [32] Friedrich Pukelsheim. *Optimal Design of Experiments*. Wiley Interscience, 1993.
- [33] Rembert Reemtsen and Jan J. Rückmann, editors. *Semi-Infinite Programming*. Springer, 1998.
- [34] Jos F. Sturm. Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12(1–4):625–653, 1999. See also <http://sedumi.ie.lehigh.edu/>.
- [35] A. F. Timan. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, Oxford, UK, 1963.
- [36] Kim Chuan Toh, Michael J. Todd, and R. H. Tütüncü. SDPT3 — a Matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11–12(1–4):545–581, 1999. doi:10.1080/10556789908805762.
- [37] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, PA, 2013.
- [38] Evgenij E. Tyrtyshnikov. How bad are Hankel matrices? *Numerische Mathematik*, 67(2):261–269, March 1994. doi:10.1007/s002110050027.
- [39] Lieven Vandenberghe and Stephen P. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, March 1996.
- [40] Liping Zhang, Soon-Yi Wu, and Marco A. López. A new exchange method for convex semi-infinite programming. *Mathematical Programming*, 20(6):2959–2977, 2010.

#### APPENDIX A. A CHARACTERIZATION OF THE SUPPORT OF OPTIMAL EXPERIMENTAL DESIGNS

The statement of the theorem requires two definitions and some notation.

First, elaborating on the definition used in Section 1, we say that the concave, continuous, objective function  $\Phi : \mathbb{S}_+^m \mapsto \mathbb{R}$  of a maximization problem is *semidefinite representable* if its closed upper level sets are semidefinite representable, that is, if for some  $k_1, \dots, k_p$  and  $\ell$  there exist linear functions  $A_i : \mathbb{S}_+^m \mapsto \mathbb{S}^{k_i}$ ,  $C_i : \mathbb{R}^\ell \mapsto \mathbb{S}^{k_i}$ , and matrices  $B_i \in \mathbb{S}^{k_i}$ ,  $D_i \in \mathbb{S}^{k_i}$  ( $i = 1, \dots, p$ ) such that for all  $X \in \mathbb{S}_+^m$  and  $z \in \mathbb{R}$ ,  $\Phi(X) \geq z$  holds if and only if there exists a  $u \in \mathbb{R}^\ell$  satisfying

$$A_i(X) + B_i z + C_i(u) + D_i \succ 0 \quad i = 1, \dots, p. \quad (16)$$

Second, we say that the semidefinite representable function  $\Phi : \mathbb{S}_+^m \mapsto \mathbb{R}$  is *admissible* with respect to the set  $\mathcal{X} \subseteq \mathbb{S}_+^m$  if  $\Phi$  has a representation (16) for which there exists an  $\hat{X} \in \mathcal{X}$  satisfying (16) with strict inequality for some  $z$  and  $u$ . In other words, the left-hand side of each of the  $p$  inequalities can be made positive definite simultaneously for at least one  $\hat{X} \in \mathcal{X}$ .

The latter is a technical condition; most interesting functions  $\Phi$  are admissible with respect to every non-empty set  $\mathcal{X}$ , or at least with respect to every  $\mathcal{X}$  that contains a non-singular matrix. Specifically, all commonly used optimality criteria, including D-, E-, and A-optimality are semidefinite representable, and they are also admissible with respect to every set of Fisher information matrices for which the criteria is well-defined.

We shall also use the common notation that the adjoint of a linear operator  $C$  is denoted by  $C^*$ .

Now we are ready to restate the characterization of the support of optimal experimental designs.

**Theorem 7** ([28]). *Suppose that in the linear model (14) the design space  $\mathcal{I}$  is a finite union of closed and bounded intervals, the functions  $f_i$  are rational functions with finite values on  $\mathcal{I}$ , and  $\omega$  is a nonnegative rational function on  $\mathcal{I}$ . Let  $\Phi$  be a semidefinite representable function with representation (16), admissible with respect to the set of Fisher information*

matrices  $\mathcal{M} = \text{conv}\{f(t)f(t)^T\omega(t) \mid t \in \mathcal{I}\}$ . Then the support of the  $\Phi$ -optimal design is a subset of the real zeros of the polynomial  $\pi$  obtained by solving the following semidefinite programming problem:

$$\begin{aligned} & \underset{\substack{y \in \mathbb{R}, \pi \in \mathbb{R}^{d+1}, \\ W_1 \in \mathbb{S}_+^{k_1}, \dots, W_p \in \mathbb{S}_+^{k_p}}}{\text{minimize}} && y \end{aligned} \tag{17a}$$

$$\text{subject to} \quad \sum_{i=1}^p \langle W_i, B_i \rangle = -1, \quad \sum_{i=1}^p C_i^*(W_i) = 0, \tag{17b}$$

$$\pi = \Pi(y, W_1, \dots, W_p), \tag{17c}$$

$$\pi(t) \geq 0 \quad \forall t \in \mathcal{I}, \tag{17d}$$

where  $d$  is the degree of the polynomial

$$t \mapsto \text{lcm}(\text{den}(\omega), \text{den}(f_1^2), \dots, \text{den}(f_m^2)) \left( y - \sum_{i=1}^p \langle W_i, A_i(M(\xi_t)) + D_i \rangle \right), \tag{18}$$

whose coefficient vector is denoted by  $\Pi(y, W_1, \dots, W_p)$  in (17c) above.

Note that the operator  $\Pi$  in (18) is affine, hence aside from (17d) every constraint in (17) is a linear equation or linear matrix inequality. Furthermore, (17d) can be translated to linear matrix inequalities using a sum-of-squares interpolant representation. Hence, (17) is indeed a semidefinite program.